DOCUMENT RESUME

ED 453 288 TM 032 827

AUTHOR Ridge, Kirk
TITLE Do Raters Demonstrate Halo Error When Scoring a Series of Responses?
PUB DATE 2001-04-13
NOTE 33p.; Paper presented at the Annual Meeting of the National Council on Measurement in Education (Seattle, WA, April 11-13, 2001).
PUB TYPE Reports - Research (143) -- Speeches/Meeting Papers (150)
EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS *Evaluators; Feedback; *Interrater Reliability; Junior High School Students; Junior High Schools; Mathematics Tests; Performance Based Assessment; Responses; *Scoring; *Training
IDENTIFIERS *Halo Effect

ABSTRACT

This study investigated whether raters in two different training groups would demonstrate halo error when each rater scored all five responses to five different mathematics performance-based items from each student. One group of 20 raters was trained by an experienced scoring director with item-specific scoring rubrics and the opportunity to practice scoring and receive feedback. A second group of 20 raters trained themselves using a generic scoring rubric without the opportunity to practice. After training, all raters scored the same 500 student responses, into which were embedded manufactured student response scenarios that were designed to induce halo error. The results indicate that the director-trained raters were more accurate than the self-trained raters on low-scoring responses, but there were no significant differences between the groups in rater halo error. A followup study was conducted using a third group of raters who were not trained and were instructed to rate the papers according to perceived quality. The results of the analysis indicated a significant difference between this group and the director-trained group in rater halo error. (Contains 2 figures, 9 tables, and 15 references.) (Author/SLD)

DO RATERS DEMONSTRATE HALO ERROR WHEN SCORING A SERIES OF RESPONSES?

Kirk Ridge
Measurement Incorporated

2001 NCME Annual Meeting

This study investigated if raters in two different training groups would demonstrate halo error when each rater scored all five responses to five different mathematics performance-based items from each student. One group of 20 raters was trained by an experienced scoring director with item-specific scoring rubrics and had an opportunity to practice scoring and receive feedback. A second group of 20 raters self-trained using a generic scoring rubric but were given no opportunity to practice. After training, all raters scored the same 500 student responses, into which were imbedded manufactured student response scenarios that were designed to induce halo error. The results indicated that the director-trained raters were more accurate than the self-trained raters on low-scoring responses, but there were no significant differences between the groups in rater halo error. A follow-up study was conducted using a third group of raters who were not trained and were instructed to rate the papers according to perceived quality. The results of this analysis indicated a significant difference between this group and the director-trained group in rater halo error.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

Introduction

Background to the Study

The major change in large-scale student assessment over the past fifteen years has been in the addition of performance assessment items. Unlike multiple-choice items, these performance, or constructed-response items, require human raters to judge student performance. As a result, additional sources of error due to raters are added to the measurement equation. The main purpose of this experiment was to study if raters in two different training conditions would demonstrate halo error when scoring responses to five different mathematics performance items from each student. This was a question believed not yet addressed by the research on the new performance assessments.

Significance

Large-scale performance assessments are a recent phenomenon. The political implications have outdistanced the research as the politicians and policy makers have moved quickly to embrace the results and impose high-stakes decisions that affect students, teachers, and schools without the new assessments being well grounded in a significant body of research. Baker (1990) cautioned, "It is also clear that the impact of tests in the service of accountability is not unbridled good" (p. 7). Messick (1989) discussed the social consequences regarding the interpretation of test scores, the relevance of the scores to the purpose, the implications for action, and the intended and unintended consequences of the test.

The move away from assessments that are exclusively multiple choice to those that include performance items has introduced additional sources of error into the measurement equation. As item writers become more ambitious to cast the performance items into real life scenarios, the construct being measured becomes more unclear. Encouraging alternate legitimate student responses brings into question how to quantify different levels of performance. Scoring rubrics attempt to respond to this but they must simultaneously be flexible enough to encompass the variety of legitimate responses but still restrictive enough to be useful as a criterion base.

Even with well-conceived performance assessment items and well-designed scoring rubrics, raters differ in backgrounds and experiences, levels of education, ideas of what constitutes successful student performance, and attitudes about assessments in general and the consequences they have on students, teachers, and schools.

Summary

The questions confronting performance assessment and rater bias surround rater accuracy and the kinds of errors raters make. More specific to this study, in many of the newer performance assessments, an individual rater scores all of an individual student's responses to a series of performance items in a single subject area. The main purpose of this experiment was to determine whether raters would demonstrate halo error when scoring responses to five different mathematics performance items for each student.

CHAPTER 2

Literature

Halo Error

The literature has suggested that in multi-dimensional rating situations, raters are capable of halo error. Wells (1907) and Thorndike (1920) noted a halo of "general merit" that influenced the ratings across rating categories. Kingsbury (1922) indicated halo as one of four types of rater error. Additional research explored the reasons behind halo error. Rugg (1922) suggested that halo error was a result of specific definitions of the rating categories. Halo caused by a rater's consideration of the consequences of the ratings was proposed by Lawler (1967). Borman (1979) related halo to a lack of precision in rating scales and a lack of training. Cooper (1981) hypothesized that raters' problems with memory may contribute to halo error, causing the rater to resort to "preexisting conceptual schemes." In a detailed review of the literature on the quality of rating data, Saal, Downey, and Lahey (1980) suggested that there are fairly consistent conceptual definitions of halo. These were identified as raters applying a global impression across different rating dimensions, inability or intransigence in distinguishing between dimensions, or tendency to use uniform ratings across dimensions.

Rater Training

Kingsbury (1922) was one of the first to stress the need for rater training, recommending a criterion-grounded training manual and a training process involving practice in using the rating instruments and getting feedback from expert trainers. Kingsbury built on the work of Rugg (1922) who called for "competent judges" to rate abstract traits and characteristics. Brown (1968), Latham et al (1975), and Bernardine and Walter (1977) conducted studies that confirmed that raters who were trained would demonstrate less halo error. McIntyre et al. (1984) and Pulakos (1984) found that training improved rater accuracy. Borman (1979) recommended additional research on training raters to learn correct performance standards that may produce more accurate ratings. He added that some dimensions may be more difficult to rate than others.

CHAPTER 3

Methods and Procedures

Introduction

In performance assessments in which a student responds to a series of items, halo is defined as a rater letting expectations about how a student should perform based on the student performance on the earlier items influence the rater's scoring of the later items. Halo error in this experiment was studied by comparing rater scores on a series of five student responses to five different mathematics performance-based items with estimated true scores established by a committee of experts. The central research question sought to determine if the duration and specificity of training would cause differences in rater halo error when scoring multiple items from an individual student. The study was further designed to determine if raters would demonstrate halo error in conditions where they were induced to score the item too high and in conditions where they were induced to score the item too low.

Raters were placed into one of two groups based on their type of training. The training groups were structured in a way that was expected to make conditions optimal to reveal group differences in halo error, based on suggestions in the literature that more specific training would reduce halo error. The first group consisted of experienced raters who received two days of training from an expert scoring director. Raters in this group used item-specific scoring rubrics and had an opportunity to practice scoring and ask questions. The second group consisted of novice raters. This group self-trained using generic scoring rubrics with no opportunity to practice or ask questions. The prediction was that there would be statistically significant differences between groups in halo error, with the director-trained group demonstrating less halo error than the self-trained group.

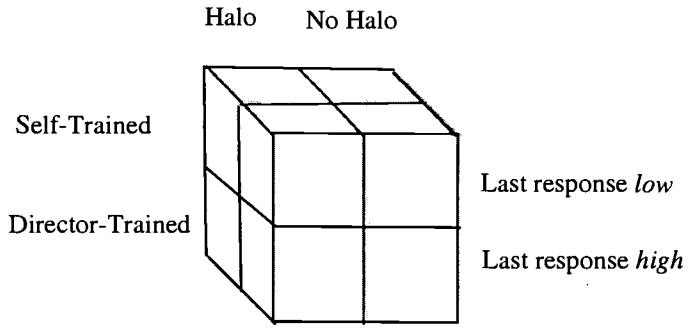
Overview

Regarding halo error, the central question sought to determine if raters would demonstrate halo error when scoring five different open-ended mathematics responses from each student. Raters were presented with two halo-inducing scoring scenarios. The first scenario was designed to determine if raters

would demonstrate halo error by scoring a *low*-scoring fifth response (scored "1") inappropriately high, when they were first presented with a series of four *high*-scoring responses (scored "3"). The second scenario was designed to determine if raters would demonstrate halo error by scoring a *high*-scoring fifth response (scored "3") inappropriately low, when they were first presented with a series of four *low*-scoring responses (scored "1"). Raters also were presented with two non-halo-inducing scoring scenarios -- a series of five *high*-scoring responses from each student and a series of five *low*-scoring responses from each student. These scenarios were included in an attempt to disentangle halo error from inaccuracy -- to determine if raters were inaccurate regardless of whether the scoring scenario was designed to induce halo error or not. The 2X2X2 experimental design is depicted in Figure 1.

Figure 1

The 2X2X2 Completely Crossed Experimental Design Used In This Study



To clarify Figure 1, one factor was the scoring scenario, whether the combination of responses was designed to elicit halo or not (halo/no halo). A second factor was group, differentiated by rater experience and also the duration and specificity of training (self-trained/director-trained). The groups were designed to promote the likelihood that the director-trained group would not halo and the self-trained group would halo. The third factor was the score of the last response (last response *low*/last response *high*), to separately analyze rater scoring tendencies at each end of the scale.

Criterion of Scoring Accuracy

Following the research of Borman (1977), the accuracy of the scores of each rater group was judged by comparing the rater's scores with the estimated true scores on the responses as determined by a committee of experts. This committee convened before the statewide scoring project in March, 1998 and was comprised of six participants: two state department curriculum and testing specialists, two classroom teachers who sat on state curriculum committees, and two of the scoring contractor's lead trainers. All of the committee members had participated in this type of proceeding before.

The student responses that were scored by the committee were selected from approximately 20 districts from across the state in an effort to be representative of the range of responses the raters would eventually encounter. The committee of experts spent approximately one day on each of the five items. As more papers were examined for a specific item, preliminary scoring decisions were reconsidered and the item-specific rubric was put into its final form. The committee reached consensus on approximately 100 student responses for each item.

The training papers and the manufactured targeted papers in this study had estimated true scores as determined by the committee of experts. All of the papers in the study were considered to be reasonably solid examples of the score points.

Tests That Were Used

Actual student responses to mathematics open-ended questions were randomly selected from an annual statewide assessment of approximately 80,000 Grade 8 students that occurs in the spring of the year. According to the design of the assessment, each student responded to five different mathematics performance items (numbered items 10, 20, 30, 40, and 50) as part of a 1 hour 50 minute mathematics assessment that also included 45 multiple choice questions and questions where students were given a space to grid their responses (similar to completion questions). The performance items required students to show their work and explain their answers in writing. In the 1998 mathematics assessment, each one of the five skill clusters of mathematics specified under the Grade 8 mathematics curriculum (Table 1) was represented by one constructed response question.

Table 1

Skill Clusters On Which The Five Performance Questions Were Based

Question #	Skill Measured
10	Data Analysis
20	Measurement/Geometry
30	Patterns/Relationships
40	Numerical Operations
50	Pre-Algebra

Scoring Scale

The scoring scale used to score all performance items in this study was a 0-3 scale, with 3 indicating the highest level of performance. Responses that received a score of "3" were for the most part accurate and demonstrated complete understanding. Responses that received a score of "2" demonstrated partial understanding and/or were incomplete. Responses with a score of "1" only began to answer the question. Responses that were irrelevant, inappropriate, or otherwise without merit received a score of "0."

Raters

Raters were selected from the pool of approximately 500 professional raters who worked for a testing company that contracts with state departments of education to score statewide performance assessments. All raters had supplied proof of a four year college degree and had undergone a lengthy interview process before being hired. Approximately 30% of the raters hired had graduate-level degrees and 20% had some teaching experience. Balance of gender, age, and ethnicity mirrored the general population.

Training and Scoring

Training of Raters

The rater training and scoring activities for this study were conducted in January 1999. There were two groups of twenty raters each, and each group was divided into two teams each. All raters in the study completed the interview process described above and were approved to be professional raters. The

raters in Group 1 were selected at random from the pool of approximately 200 experienced raters who were available at the time of the study. The raters in Group 2 were selected at random from the pool of approximately 50 inexperienced raters who were available at the time of the study.

Group 1 was trained following the same 2-day training protocol that was followed during the actual test scoring. Group 2 received abbreviated training that lasted 1/2 day. Raters in both groups were informed that they were part of a scoring project that was involved with researching different training methods, but they were not informed of any of the specifics of the study. All raters were told that they were to do their best to score the papers according to the scoring protocol covered in training and that it was essential, as in all scoring projects, to score according to the state's criteria and to stay free of personal bias.

Group 1 was trained by the same scoring director who trained raters during the spring 1998 scoring project. She had conducted similar training for over 20 state assessment projects, at least one half of which were mathematics assessments. In addition, the scoring director was a member of the committee of experts that determined the scores on all of the papers used in this study.

The training of Group 1 was conducted according to the best practices in the field today. This included item-specific scoring rubrics that were criterion-based to delineate clearly the gradients of student performance. Raters were trained by an expert scoring director who carefully explained the rationale behind the scores on the training papers. Further, the scoring director taught the raters to use both the item-specific scoring rubrics and the anchor papers as tools that together defined the scoring scale. The scoring director also discussed with raters how to determine scores on the range and variety of responses within each score point. The process of practicing scoring and receiving guidance and feedback from the scoring director was designed to help the raters in Group 1 refine their scoring skills during the two days of training and develop confidence in using the scoring rubrics and anchor papers.

Group 2, consisting of 20 raters, was self trained. Although these were inexperienced raters, all 20 had four-year college degrees and ten had teaching experience. In many assessments raters are selected based on their experience in the field. They are often given little additional training because the assumption is made that because they work in the field, they have the related skills needed to evaluate performance. Raters in Group 2 received a generic mathematics scoring

rubric and the same 80 anchor papers as did raters in Group 1. Although Group 2 had a scoring director to offer encouragement and to facilitate paper flow, no discussion or clarification of the rationale of the scores on the papers occurred. Because of the lack of discussion, the Group 2 training was shorter in duration than the Group 1 training.

Scoring Procedure

Student papers for the study were randomly-selected responses from 100 Grade 8 students. Student responses were assembled into five packets of 20 students each, with five responses from each student (questions #10, 20, 30, 40, and 50). Two of the twenty students in each packet were manufactured targeted student scenarios designed to elicit halo error. Across five packets, then, raters scored ten manufactured targeted responses. The paper design was as designated in Table 2.

Table 2

Number of Targeted and Untargeted Students Scored by Each Rater

Total # of Students in a packet	20
Total # of Responses in a packet	100 (five items per student)
# of untargeted students in the packet	18 (same 18 students scored by Team A and Team B)
# of targeted students in a packet	2
Total # of packets scored by each rater	5

Five packets of 20 students (five responses each, one to each of the five items) were randomly selected from the 4,000 packets of actual student responses from the 1998 test. Eighteen of the students in each of the five packets were kept in their original place and scored by all raters, regardless of group or team. Two of the students in each packet were replaced with manufactured targeted students, the responses of which were selected from the responses scored by the committee of experts. Each of the manufactured targeted student responses was copied over onto actual blank student booklets in the same handwriting to make it appear as if they were written by the same student. These targeted responses were then placed into the packets of responses so they were indistinguishable from the other responses in the packet.

The targeted responses were assembled into four different types of scenarios. Two of the scenarios were designed to induce halo error -- either four *high*-scoring responses (scored "3") followed by a *low*–scoring response (scored "1") or four *low*-scoring responses (scored "1") followed by a *high*-scoring

response (scored "3"). Two of the scenarios were designed to determine accuracy of scoring a series of responses that was not designed to induce halo error -- either five *low*–scoring responses, or five *high*-scoring responses. No responses that scored "0" were used as targeted responses because the score of "0" is reserved for highly unresponsive papers that are either so short or so lacking in focus that they would be too obvious to the raters to be of use in this halo study.

Forty raters were assigned to either the director-trained or the self-trained group. Each group of 20 raters was divided into two teams, Team A and Team B, of ten raters each. Eighteen of the student responses in each packet were identical for each team. Two of the student responses in each packet were different manufactured targeted student scenarios for each team.

The targeted responses were arranged into scenarios that were designed to optimize the opportunity to produce halo error. In each packet, raters in Team A encountered one student who had four *low*-scoring responses followed by a *high*-scoring response (halo student) and one student who had four *low*-scoring responses followed by a *low*-scoring response (non-halo student), as depicted in Table 3.

Table 3

Pattern of Manufactured Targeted Halo Students for Team A

Item #	10	20	30	40	50
Team A - Target Halo Student #1a	1	1	1	1	3
Team A - Target Non-Halo Student #1b	1	1	1	1	1

(Items 10, 20, 30, and 40 were the same responses for Student #1a and #1b. The responses to #50 were different. All five responses for each student were in the same handwriting)

Raters in Team B encountered one student who had four *high*-scoring responses followed by a *low*-scoring response (halo student) and one student who had four *high*-scoring responses followed by a *high*-scoring response (non-halo student) in each of the five packets, as depicted in Table 4.

Table 4

Pattern of Manufactured Targeted Halo Students for Team B

Item#	10	20	30	40	50
Team B - Target Halo Student #1a	3	3	3	3	1
Team B – Target Non-Halo Student #1b	3	3	3	3	3

(Items 10, 20, 30, and 40 were the same responses for Student #1a and #1b. The responses to #50 were different. All five responses for each student were in the same handwriting)

Summary of Procedures Used

Every effort was made to exercise as much control as possible in this study. Actual student responses used in the training and scoring processes were selected at random from the packets of responses from the 1998 Grade 8 Administration. The scores on the training papers and on the targeted student responses were pre-established by a committee of experts. The scoring director followed a well-established training protocol. Paper flow was orchestrated according to a crossed design, and each of the manufactured targeted student scenarios was copied in the same handwriting into actual blank student booklets and spiraled into the packets so that they were indistinguishable from the other responses in the packets. The numbers of raters in each group were equal and no raters dropped out of the study. Regardless of group, after training the scoring director did not answer any scoring-related questions. Raters in both groups took approximately 21/2 days to score the five packets of student responses.

CHAPTER 4

Findings

Overview

The first, and central analysis was conducted to determine if the two training groups differed in demonstrating halo error as evidenced by a rater scoring a *low*-scoring response to Item #50 inappropriately high, if it followed a preceding series of four *high*-scoring responses (on Items 10-40), or by scoring a *high*-scoring response to Item #50 inappropriately low, if it followed a preceding series of four *low*-scoring responses. The analysis found no statistically significant main effect for halo.

Analysis of Halo Error in Targeted Responses (#50)

All statistical analyses were evaluated at an alpha level of .05. The 2X2X2 factorial design was analyzed using SPSS for Windows 6.1.2. The analysis was conducted to determine main effects for Group – trained by the scoring director vs. self-trained, Halo – whether the targeted responses contained a scenario that was designed to encourage halo (33331 / 11113) or a scenario that was not designed to encourage halo (33333 / 11111), and Score of Item 50 – whether the last response was a 1 (33331 / 11111) or a 3 (33333 / 11113).

The dependent variable was calculated by first determining the absolute value of the difference between each rater's score and the expert score on each of the Item #50's of the targeted responses. These values were then summed across the five packets then divided by five (the number of packets) to produce the average absolute difference from the expert scores on Item #50 for each rater on all of the items across the five packets. Table 5 shows the results of the 2X2X2 ANOVA.

Table 5

Results of the 2X2X2 ANOVA

Source of Variation	*MS*	*DF*	*F*	*Sig.*
Main Effects				
GROUP	1.98	1	32.33	<.001
SCORE OF ITEM #50	1.40	1	22.88	<.001
HALO	.18	1	2.94	.09
2-Way Interactions				
GROUP/SCORE OF #50	.68	1	11.15	<.001
GROUP/HALO	.11	1	1.83	.18
SCORE OF #50/HALO	.04	1	.66	.42
3-Way Interactions				
GROUP/#50/HALO	.01	1	.20	.65
Residual	.06	72		

As indicated in Table 5, the ANOVA revealed statistically significant main effects ($p \leq .05$) for Group, $F(1, 72) = 32.33$, $p<.001$ and Score of Item 50, $F(1, 72) = 22.88$, $p<.001$ but not for Halo Scenario, $F(1, 72) = 2.94$, $p=.09$. The outcome of the main effect for Group indicated that the groups differed in their scoring accuracy, as was expected because of the difference between groups in their duration and specificity of training. As can be seen in the means reported in Table 6, the director-trained group deviated on average from the pre-established score by only .04 points while the self-trained groups deviated on average from the pre-established score by .36 points.

The statistically significant main effect for Score of Item 50 was not predicted, but it was noted as a possibility when the study was designed. As indicated earlier in this paper, the Score of Item 50 factor was included to enable separate analyses of rater scoring of Item #50 responses with an expert score of "1" and Item #50 responses with an expert score of "3." As can be seen in the means reported in Table 6, the deviation in the scores on Item 50 designated as low scoring (score of "1") from the pre-established score

was much greater (.33) than the deviation of high score items (score of "3") from the pre-established score (.07).

Also not predicted was the outcome for the Halo Scenario factor. It was expected that scores of Item #50 responses that were in a halo scenario would be less accurate than scores in a non-halo scenario, thus revealing halo. This was not the case, as the analysis showed no statistically significant difference between Item #50 responses in the halo scenario (.25) versus the non-halo scenario (.15).

Table 6

Cell Means and Standard Deviations for Group, Halo Scenario, and Score of Item 50

	Group	
	Director-Trained	Self-Trained
M	.04	.36
SD	.11	.40
	Halo Scenario	
	Yes	No
M	.25	.15
SD	.40	.24
	Score of Item 50	
	1	3
M	.33	.07
SD	.35	.26

The 2X2X2 analysis partitioned variation due to differences between groups into each pair of independent variables as well as the main effects. The ANOVA revealed a statistically significant two-way interaction ($p \leq .05$) for Group and Score of Item 50, $F(1,72) = 11.15$, $p=.01$ but not for Group and Halo Scenario, $F(1,72) = 1.83$, $p=.18$ or for Score of Item 50 and Halo Scenario, $F(1,72) = .66$, $p=.42$. Table 7 shows the cell means and standard deviations for the two-way interaction of Group and Score of Item 50.

Table 7

Cell Means and Standard Deviations for Interaction of Group and Score of Item 50

Group		Score of Item 50	
		1	3
Director-Trained	*M*	.08	.00
	SD	.15	.00
Self-Trained	*M*	.58	.13
	SD	.31	.36

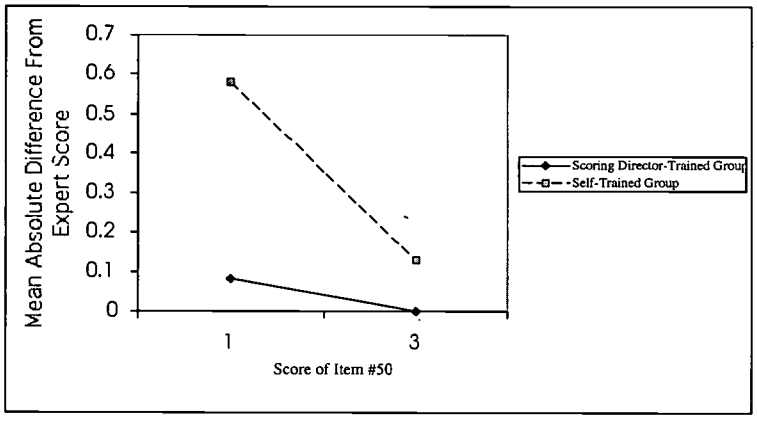


Figure 2. Graph of the 2-way interaction of Group and Score of Item 50.

As Table 7 and the graph in Figure 2 indicate, the means of the two training groups differed considerably more on the Item #50 responses with an expert score of "1" than they did on the Item #50 responses with an expert score of "3." As indicated by the group means, the two training groups only differed in their scoring of Item #50 responses with an expert score of "3" by .13, but they differed in their scoring of Item #50 responses with an expert score of "1" by .50. This resulted in the statistically significant two-way interaction of Group and Score of Item 50.

Secondary Analysis of Accuracy of Scores on Items 10, 20, 30, 40, and 50 of Targeted Students

In addition to the primary analysis of the scores of Item #50 that was conducted to determine halo error, a secondary analysis was conducted to determine the accuracy of scores on all five items for the targeted students. This was considered important in determining if the scoring patterns of each training group on Item #50 differed from their scoring patterns across all five items.

The dependent variable was calculated by first determining the absolute value of the difference between each rater's score and the expert score on each of the five items (10, 20, 30, 40, 50) of the targeted students. These values were then summed across the five packets then divided by 25 for the individual analyses of all items/score point "1" and all items/score point "3" to produce the average absolute difference from the expert scores for each rater on all of the items across the five packets.

Regarding scoring accuracy, a 2x2x2 ANOVA was calculated across all five items (10, 20, 30, 40, 50) for the targeted students. This analysis was conducted to determine if the training groups scoring patterns across all items were consistent with their scoring of Item #50. The results of the 2x2x2 ANOVA revealed statistically significant main effects ($p \leq .05$) for Group, $F(1, 36) = 56.85$, $p<.001$ and Score of #50, $F(1, 36) = 56.85$, $p<.001$. The two-way interaction of Group/Score of #50 was also significant $F(1, 36) = 5.95$, $p=.02$. The results were consistent with the findings of the scores on Item #50. The training groups showed significant differences between groups on the scores across all items with a pre-established score of "1" but showed no significant differences across all items with a pre-established score of "3".

Summary

The 2X2X2 factorial analysis of Item #50 revealed no statistically significant main effect ($p \leq .05$) for Halo but did indicate statistically significant main effects ($p \leq .05$) for Group, $F(1,72) = 32.33$, $p < .001$ and Score of Item 50, $F(1,72) = 22.88$, $p <.001$. The two-way interaction indicated a group difference for responses that had an expert score of "1", regardless of whether these were in a halo-promoting scenario (33331) or not (11111). On these, the self-trained group was significantly less accurate than the director-trained group, $F(1,72) = 11.15$, $p < .001$. The results of the scoring across all five items were consistent with the results of the scoring of Item #50.

CHAPTER 5

Conclusions and Discussion

The Central Question - Halo Error

Did the two training groups differ in demonstrating a halo produced by a series of good or poor responses? This study was not able to show group differences in halo. This does not mean that halo error does not happen, but that it did not happen under the conditions of this experiment.

The results of this experiment did not reveal a statistically significant main effect for Halo Scenario, $F(1, 72) = 2.94$, $p=.09$. In addition, there was not a statistically significant 2-way interaction for Halo and Group $F(1, 72) = 1.83$, $p=.18$ or for Halo and Score of Item 50, $F(1,72) = .66$, $p=.42$ nor was there a significant 3-way interaction of Halo by Group by Score of Item 50 $F(1,72) = .20$, $p=.65$. This outcome was not predicted. It was expected that the training groups would show significant differences in halo error.

Group Main Effect and Group by Score of Item 50 Interaction

This study was designed to apply the findings of the previous studies of halo error in a new setting – a mathematics performance assessment. Unlike prior studies that analyzed average standard deviations or interdimensional correlations to ascertain halo error, this study manufactured student scenarios that were designed to specifically elicit halo error and imbed them in a larger sample of student responses. Groups of raters were trained according to two different levels of training specificity and duration, with the assumption that the training group that self-trained would demonstrate halo error significantly more than the group that was trained by an experienced scoring director with item-specific scoring rubrics and an opportunity to practice and receive feedback.

Rugg (1922) and Kingsbury (1922) hypothesized that rater training would reduce halo error. This was confirmed in studies by Brown (1968), Latham et al (1975), and Bernardine and Walter (1977). None of these studies, however, indicated the specificity and duration of training required to reduce halo error to an insignificant level.

The self-trained group in this study did have a certain level of training. The self-trained group received the same anchor papers as did the director-trained group and they also received a generic scoring rubric that gave general descriptions of the scoring scale. Also, all of the self-trained raters had four year college degrees and were cleared through the testing company's interview process. The self-trained raters, like the director-trained raters, believed that this study was like all scoring projects in that they were to attempt to apply the scoring criteria as accurately as possible, based on the training that they received. It is possible that both groups of raters in this study were too well trained to fall into the halo trap as defined by the manufactured student responses.

The two-way interaction did indicate that the self-trained raters were significantly less accurate than the director-trained raters on the Item #50 responses both in the 33331 and the 11111 scenarios. That the groups differed in scoring the Item #50 responses that were "1's" and did not differ in scoring the Item #50 "3's" may be due in part to the fact that the responses with an expert score of "3" were easier calls. The "1" score point was not at the extreme end of the scale, unlike the "3" score point. Thus, a rater incorrectly scoring a paper with an expert score of "3" had only one way to go – down. A rater incorrectly scoring a paper with an expert score of "1" could miss it on either side, by giving a "0" or a "2." In other words, although the groups did not differ with respect to halo or in scoring the Item #50 responses that were "3's," the more specific training received by the director-trained group may have enabled them to more accurately make the scoring discriminations on the Item #50 responses with an expert score of "1".

The two-way interaction in this study may suggest that different levels of the duration and specificity of rater training may be needed to reduce halo error and increase accuracy. Abbreviated but still criterion-grounded training such as received by the self-trained group in this study may be sufficient to reduce halo error. More sustained and specific training, such as that received by the director-trained group in this study, may be needed to improve rater accuracy.

Of note is that the self-trained raters had little experience in scoring student responses. The raters in the director-trained group were more experienced, having worked in several prior scoring projects. An analysis of data from the 1998 statewide scoring of these items, however, revealed no significant difference between experienced and novice raters in rating accuracy after they went through identical in-depth training similar to that received by the director-trained raters in this study. Therefore, it is likely that item-specific

scoring rubrics, training by an experienced scoring director, and having the opportunity to practice scoring and receive feedback on correct and incorrect scores may have a significant effect on rater accuracy. Additional research is recommended to confirm these findings, including additional investigations into whether improving the accuracy of ratings is due more to training than to scoring experience.

The results of this study should not be generalized to other scoring situations such as projects in which raters score using scales with more score points or in which they score more than five responses to different items from each student. The results should not be generalized to the scoring of other types of performance assessment items. By nature, the scoring of mathematics performance items is more objective and rule driven than the scoring criteria for language arts items. Additional research is recommended into whether raters demonstrate halo error and/or lowering of accuracy when they score multiple items in reading, writing, science, or social science and when they score single items that are scored on multiple dimensions, as is the case in a increasing number of assessments.

Recommendations for future research also include examining how many items an individual rater can score before accuracy is compromised. For example, raters may be able to score five items on one occasion accurately, but begin to lose accuracy if they have to score ten on one occasion. Some of the issues here relate to how much scoring information a rater can hold in short term memory and at what point the corresponding overload leads to cognitive dissonance.

Summary

There is increased emphasis on using the scores from performance assessments to make high-stakes decisions about individual students. The results of this study suggest that halo error does not appear to be a significant issue at the group level, but rater accuracy is still a concern. Additional experiments that study halo error and accuracy in other multidimensional rating situations are recommended. This study also revealed indications that rater training methodology may have an effect on scoring accuracy. Research is needed to confirm the significance of training designs that include an expert trainer, more specific training materials, and longer training with the opportunity to practice. The literature discusses the types of rater errors but seldom covers the reasons behind the errors and the extent to which training can isolate and address these reasons for these errors. The two-way interaction in this study may suggest that different levels of the duration and specificity of rater training may be needed to reduce halo error and increase

accuracy. Abbreviated but still criterion-grounded training such as received by the self-trained group in this study may be sufficient to reduce halo error. More sustained and specific training, such as that received by the director-trained group in this study, may be needed to improve rater accuracy.

CHAPTER 6

Follow-up Study

Design

The results of the initial study suggested that abbreviated but still criterion-grounded training such as received by the self-trained group was sufficient to reduce halo error. In an attempt to explore this hypothesis further, a third group of 20 raters was assembled to score the same papers as scored by the initial two groups of raters. This third group was comprised of experienced raters of a similar profile to the raters in Group 1. Group 3, however, received no training. They were given the five math items and no other materials. They were instructed to spend a few minutes becoming familiar with the items and then to score all of the papers on a 0-3 scale, relying on their sense of the quality of the responses as they essentially rank ordered the papers.

The assumption was that, with a lack of even the minimal training materials given to the self-trained group in the initial study, this third group would rely on their own beliefs of implicit covariance between a student's performance across the five items, as suggested by the research on halo error. In other words, lacking any criterion-based rating materials, the raters in Group 3 would be influenced by evidence across the first four items in scoring the fifth, consequently demonstrating halo error. The scores of the third group were compared with the scores of the first group, as presented in Table 8.

Table 8

Results of the 2X2X2 ANOVA

Source of Variation	*MS*	*DF*	*F*	*Sig.*
Main Effects				
GROUP	1.30	1	29.15	<.001
SCORE OF ITEM #50	.61	1	13.73	<.001
HALO	.26	1	5.93	.02
2-Way Interactions				
GROUP/SCORE OF #50	.14	1	3.24	.08
GROUP/HALO	.22	1	4.94	.03
SCORE OF #50/HALO	.04	1	.91	.34
3-Way Interactions				
GROUP/#50/HALO	.03	1	.55	.46
Residual	.06	72		

As indicated in Table 8, the ANOVA revealed statistically significant main effects ($p \leq .05$) for Group, $F(1, 72) = 29.15$, $p<.001$, Score of Item 50, $F(1, 72) = 13.73$, $p<.001$, and Halo Scenario, $F(1, 72) = 5.92$, $p=.02$.

The 2X2X2 analysis partitioned variation due to differences between groups into each pair of independent variables as well as the main effects. The ANOVA revealed a statistically significant two-way interaction ($p \leq .05$) for Group and Halo Scenario, $F(1,72) = 4.94$, $p=.03$ but not for Group and Score of Item 50, $F(1,72) = 3.24$, $p=.08$ or for Score of Item 50 and Halo Scenario, $F(1,72) = .91$, $p=.34$. Table 9 shows the cell means and standard deviations for the two-way interaction of Group and Halo Scenario.

Table 9

Cell Means and Standard Deviations for Interaction of Group and Halo Scenario

Group		Halo Scenario	
		Yes	No
Director-Trained	*M*	.05	.04
	SD	.15	.00
No Training	*M*	.41	.19
	SD	.41	.14

As Table 9 indicates, the means of the two groups differed considerably more on the Item #50 responses in the halo scenario than they did on the Item #50 responses in the non-halo scenario. As indicated by the group means, the two training groups only differed in their scoring of Item #50 responses in the non-halo scenario by .15, but they differed in their scoring of Item #50 responses in the halo scenario by .36. This resulted in the statistically significant two-way interaction of Group and Halo Scenario.

Conclusions

This study was completed shortly before the 2001 NCME Annual Meeting, and the author has not yet completed a thorough analysis of the results. Nevertheless, one possible explanation is that due to a lack of criterion-based training, the third group developed their own scoring criteria based upon the evidence at hand. When scoring five responses from an individual student, the evidence about the student based on the quality of the first four items influenced the raters in the scoring of the fifth item. If the first four responses all appeared to be strong (3333), they incorrectly scored the final response higher than the "true" score of "1" because they assumed that this was a high-performing student. If the first four responses all appeared to be weak (1111), they incorrectly scored the final response lower than the "true" score of "3" because they assumed that this was a low-performing student.

If this assumption is confirmed upon further deliberation regarding the meaning of the results, the implications are considerable. It may be that a strong training design with an expert scoring director, anchor papers, and item-specific rubrics results in a high level of scoring accuracy. Modest training with anchor papers but no item-specific rubrics or explanation by a scoring director may be sufficient to eliminate halo error but inadequate to result in a high level of scoring accuracy. With no training, and

relying on their implicit assumptions of covariance, raters may demonstrate both inaccurate ratings and halo error.

References

Baker, E.L. (1990). Developing comprehensive assessments of higher order thinking. In G. Kulm (Ed.), Assessing Higher Order Thinking in Mathematics. Washington D.C.: American Association for the Advancement of Science.

Bernardin, H.J. & Walter, C.S. (1977). Effects of rater training and diary-keeping on psychometric error in ratings. Journal of Applied Psychology, 62, 64-69.

Borman, W.C. (1977). Consistency of rating accuracy and rating errors in the judgement of human performance. Organizational Behavior and Human Performance. 20, 238-252.

Brown, E.M. (1968). Influence of training, method, and relationship on the halo effect. Journal of Applied Psychology, 52, 195-199.

Cooper, W.H. (1981). Ubiquitous halo. Psychological Bulletin, 90, 218-244.

Kingsbury, F.A. (1922). Analyzing ratings and training raters. Journal of Personnel Research, 1, 377-383.

Latham, G.P., Wexley, K.N., & Pursell, E.D. (1975). Training managers to minimize Rating errors in the observation of behavior. Journal of Applied Psychology, 60, 550-555.

Lawler, E.E. (1967). The multitrait-multirater approach to measuring managerial job performance. Journal of Applied Psychology, 51, 369-381.

McIntyre, R.M., Smith, D.E., & Hassett, C.E. (1984). Accuracy of performance ratings as affected by training and perceived purpose of rating. Journal of Applied Psychology, 69, 147-156.

Messick, S. (1989). Meaning and values in test validation: The science and ethics of assessment. Educational Researcher, 18, 5-11.

Pulakos, E.D. (1984). A comparison of rater training programs: Error training and accuracy training. Journal of Applied Psychology, 69, 581-588.

Rugg, H. (1922). Is the rating of human character practical? Journal of Educational Psychology, 13, 30-42;81-93.

Saal, F.E., Downey, R.G., & Lahey, M.A. (1980). Rating the ratings: Assessing the psychometric quality of rating data. Psychological Bulletin, 88, 413-238.

Thorndike, E.L. (1920). A constant error in psychological ratings. Journal of Applied Psychology, 4, 25-29.

Wells, F.L. (1907). A statistical study of literary merit. Archives of Psychology, 7, 5-30.

TM032827

U.S. Department of Education
Office of Educational Research and Improvement (OERI)
National Library of Education (NLE)
Educational Resources Information Center (ERIC)

REPRODUCTION RELEASE

(Specific Document)

I. DOCUMENT IDENTIFICATION:

Title: Do Raters Demonstrate Halo Error When Scoring A Series of Responses?	
Author(s): Kirk Ridge, Ph.D.	
Corporate Source: Measurement Incorporated	Publication Date: April 13, 2001

II. REPRODUCTION RELEASE:

In order to disseminate as widely as possible timely and significant materials of interest to the educational community, documents announced in the monthly abstract journal of the ERIC system, *Resources in Education* (RIE), are usually made available to users in microfiche, reproduced paper copy, and electronic media, and sold through the ERIC Document Reproduction Service (EDRS). Credit is given to the source of each document, and, if reproduction release is granted, one of the following notices is affixed to the document.

If permission is granted to reproduce and disseminate the identified document, please CHECK ONE of the following three options and sign at the bottom of the page.

The sample sticker shown below will be affixed to all Level 1 documents

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL HAS BEEN GRANTED BY

Sample

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

1

Level 1

☑

Check here for Level 1 release, permitting reproduction and dissemination in microfiche or other ERIC archival media (e.g., electronic) *and* paper copy.

The sample sticker shown below will be affixed to all Level 2A documents

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL IN MICROFICHE, AND IN ELECTRONIC MEDIA FOR ERIC COLLECTION SUBSCRIBERS ONLY, HAS BEEN GRANTED BY

Sample

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

2A

Level 2A

☐

Check here for Level 2A release, permitting reproduction and dissemination in microfiche and in electronic media for ERIC archival collection subscribers only

The sample sticker shown below will be affixed to all Level 2B documents

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL IN MICROFICHE ONLY HAS BEEN GRANTED BY

Sample

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

2B

Level 2B

☐

Check here for Level 2B release, permitting reproduction and dissemination in microfiche only

Documents will be processed as indicated provided reproduction quality permits.
If permission to reproduce is granted, but no box is checked, documents will be processed at Level 1.

I hereby grant to the Educational Resources Information Center (ERIC) nonexclusive permission to reproduce and disseminate this document as indicated above. Reproduction from the ERIC microfiche or electronic media by persons other than ERIC employees and its system contractors requires permission from the copyright holder. Exception is made for non-profit reproduction by libraries and other service agencies to satisfy information needs of educators in response to discrete inquiries.

Sign here, → please

Signature: Kirk Ridge	Printed Name/Position/Title: Kirk Ridge, Ph.D. / Vice President
Organization/Address: Measurement Incorporated, 423 Morris Street, Durham, NC 27701	Telephone: (919) 683-2413 FAX: (919) 425-7733
	E-Mail Address: kridge@measinc.com Date: 4/30/01

III. DOCUMENT AVAILABILITY INFORMATION (FROM NON-ERIC SOURCE):

If permission to reproduce is not granted to ERIC, or, if you wish ERIC to cite the availability of the document from another source, please provide the following information regarding the availability of the document. (ERIC will not announce a document unless it is publicly available, and a dependable source can be specified. Contributors should also be aware that ERIC selection criteria are significantly more stringent for documents that cannot be made available through EDRS.)

Publisher/Distributor:
Address:
Price:

IV. REFERRAL OF ERIC TO COPYRIGHT/REPRODUCTION RIGHTS HOLDER:

If the right to grant this reproduction release is held by someone other than the addressee, please provide the appropriate name and address:

Name:
Address:

V. WHERE TO SEND THIS FORM:

Send this form to the following ERIC Clearinghouse:

University of Maryland
ERIC Clearinghouse on Assessment and Evaluation
1129 Shriver Laboratory
College Park, MD 20742
Attn: Acquisitions

However, if solicited by the ERIC Facility, or if making an unsolicited contribution to ERIC, return this form (and the document being contributed) to:

ERIC Processing and Reference Facility
1100 West Street, 2nd Floor
Laurel, Maryland 20707-3598

Telephone: 301-497-4080
Toll Free: 800-799-3742
FAX: 301-953-0263
e-mail: ericfac@inet.ed.gov
WWW: http://ericfac.piccard.csc.com

EFF-088 (Rev. 9/97)